Accelerating Understanding Summit 2016

Moving from compute-centric to data-centric and network-centric – the implications for HPC

Dr Martin Hilgeman, HPC Consultant EMEA - Dell
Dr Thomas Connor, Senior Lecturer - Cardiff University
Dr Herbert Cornelius, Technical Director Advanced Computing EMEA, Intel

HPC: data centric or compute centric?

Martin Hilgeman

HPC Consultant EMEA

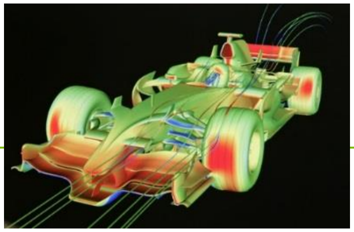
HPC Is Transforming

Traditional High Performance Computing

Computationally-intensive modeling & simulation applications by scientists, engineers and others

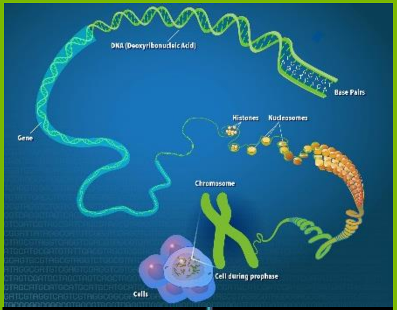
Traditional modeling and simulation applications:

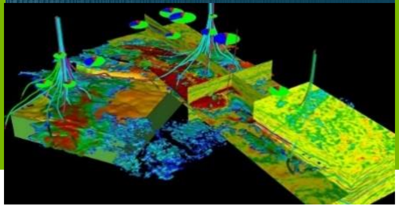
- Computer-aided design and manufacturing (CAD/CAM/CAE)
- Weather forecasting
- Oil Exploration



Data-centric HPC applications

- Genomics
- Seismic analysis
- Signal processing





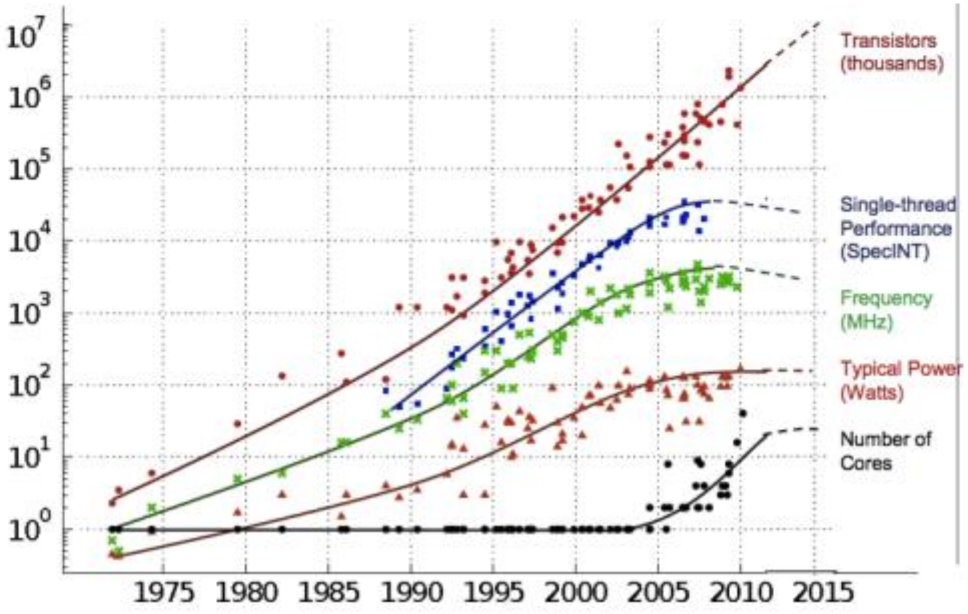
High Performance Data Analytics (HPDA)

Using HPC technologies to analyze big data for rapid insights, real time results and predictive analytics

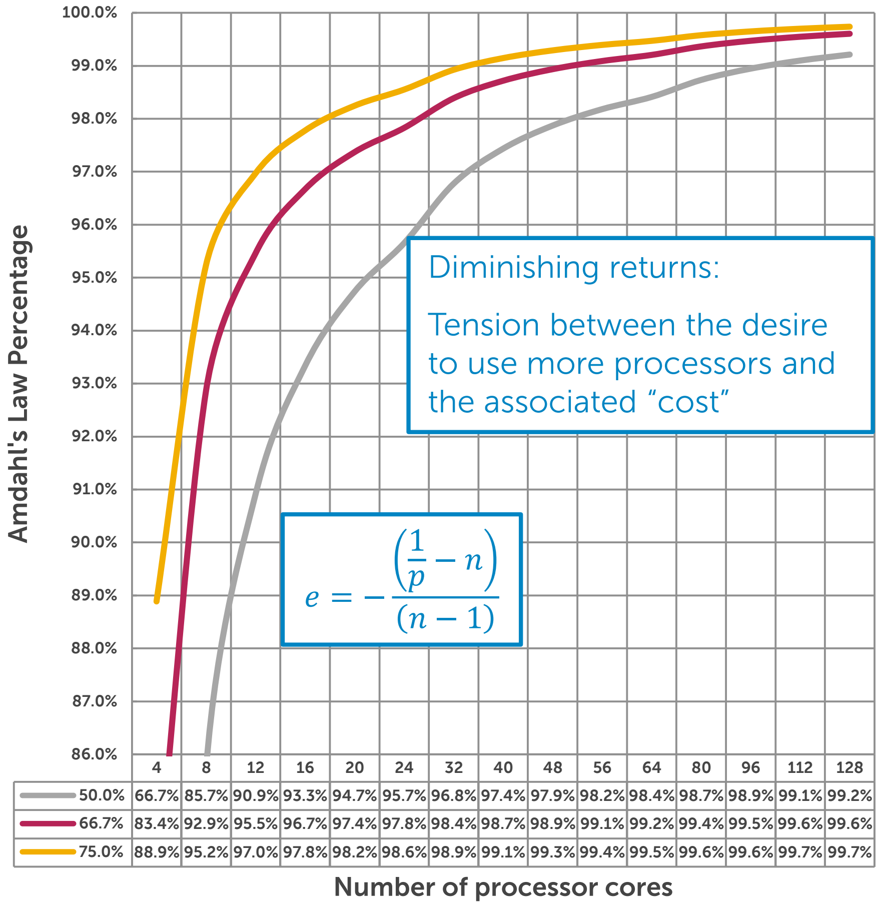
New HPDA applications:

- Personalized medicine
- Fraud detection
- Marketing

Moore's Law vs. Amdahl's Law



Chuck Moore, "DATA PROCESSING IN EXASCALE-CLASS COMPUTER SYSTEMS", The Salishan Conference on High Speed Computing, 2011



Diminishing returns:

Tension between the desire to use more processors and the associated "cost"

$$e = -\frac{\left(\frac{1}{p} - n\right)}{(n-1)}$$

	4	8	12	16	20	24	32	40	48	56	64	80	96	112	128
50.0%	66.7%	85.7%	90.9%	93.3%	94.7%	95.7%	96.8%	97.4%	97.9%	98.2%	98.4%	98.7%	98.9%	99.1%	99.2%
66.7%	83.4%	92.9%	95.5%	96.7%	97.4%	97.8%	98.4%	98.7%	98.9%	99.1%	99.2%	99.4%	99.5%	99.6%	99.6%
75.0%	88.9%	95.2%	97.0%	97.8%	98.2%	98.6%	98.9%	99.1%	99.3%	99.4%	99.5%	99.6%	99.6%	99.7%	99.7%

- The clock speed plateau
- The power ceiling
- IPC limit

- Industry is applying Moore's Law by adding more cores
- Meanwhile Amdahl's Law says that you cannot use them all efficiently

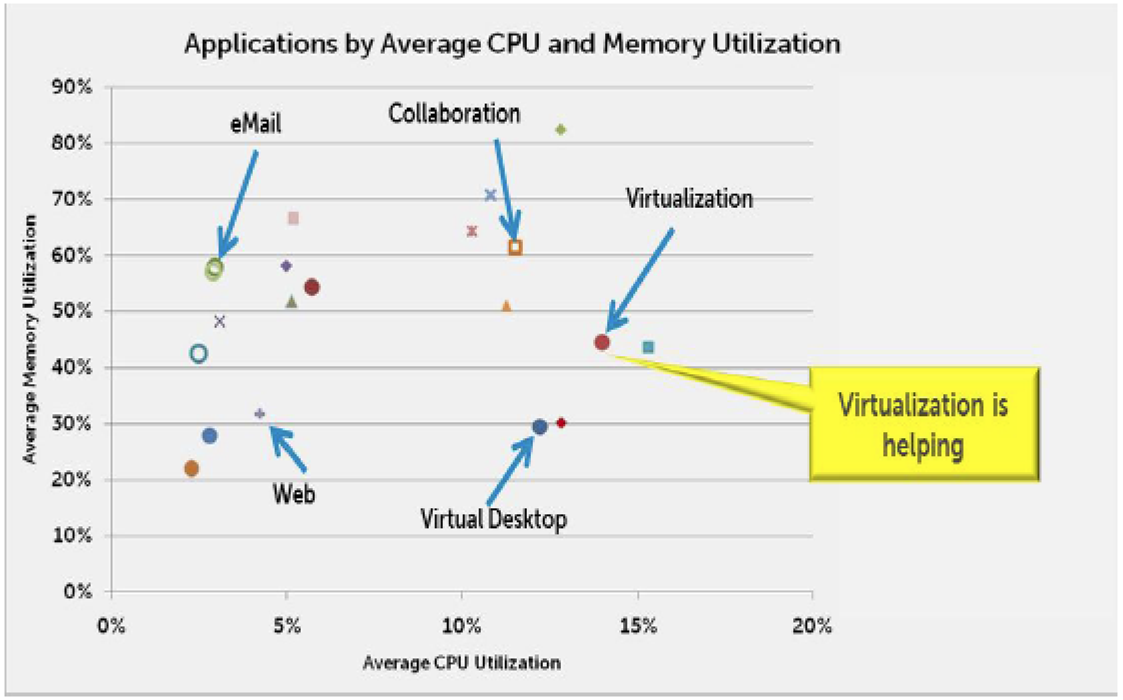
Moore's Law vs Amdahl's Law - "too Many Cooks in the Kitchen"

Industry is applying Moore's Law by adding more cores

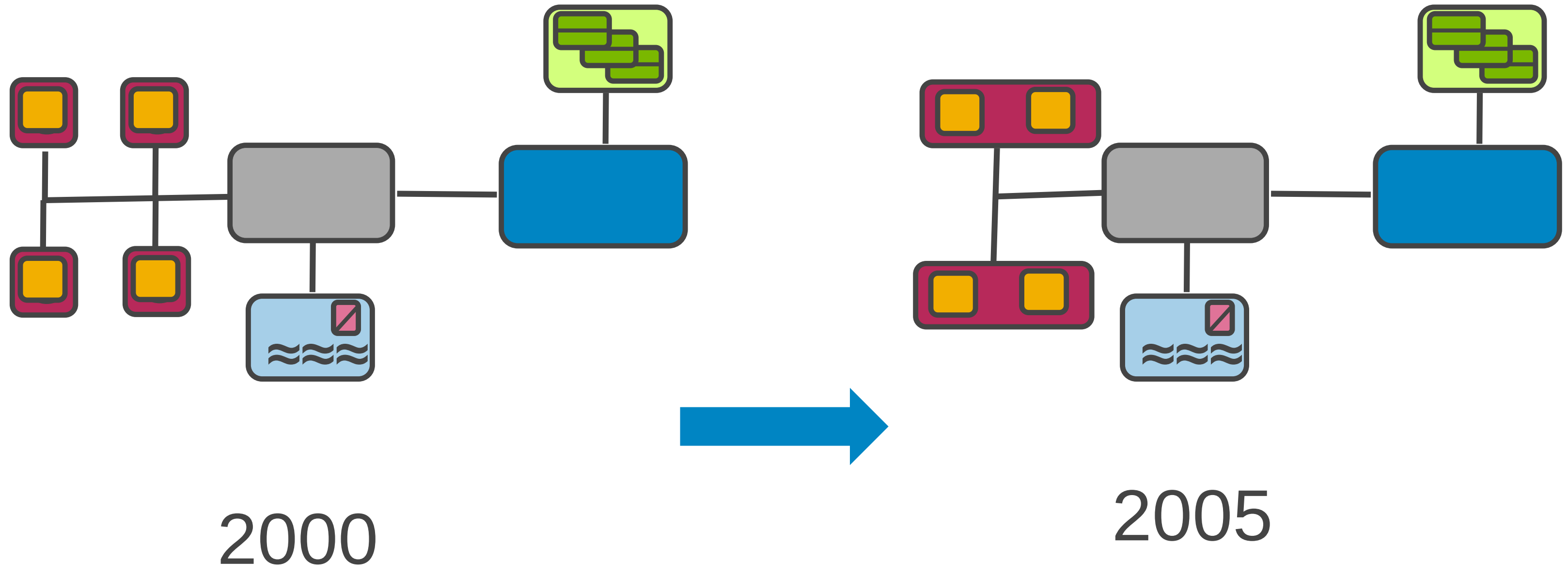
Meanwhile Amdahl's Law says that you cannot use them all efficiently

Meanwhile… traditional IT is swimming in performance

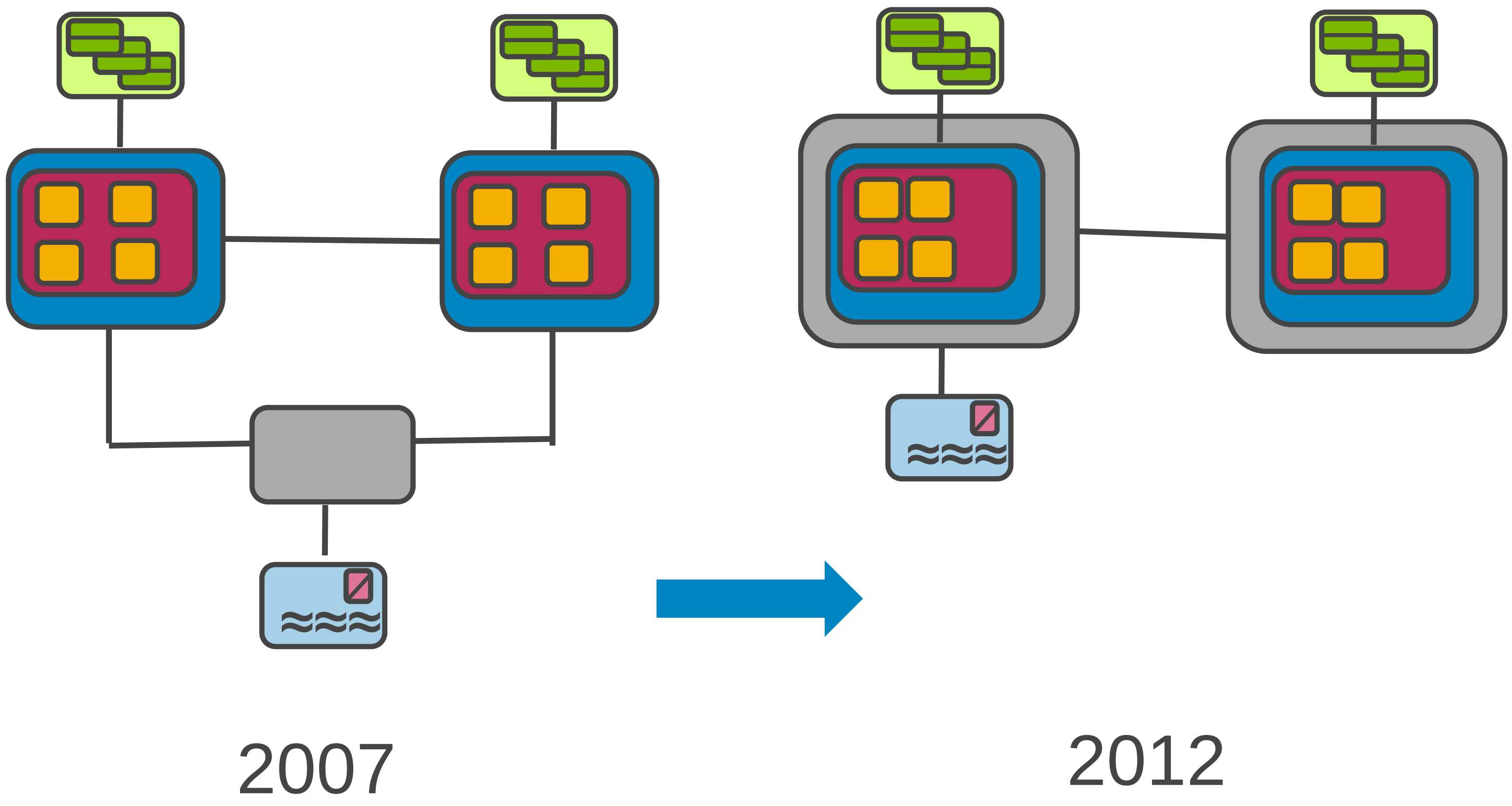
- Traditional IT server utilization rates remain low
- New µServers are emerging, x86 and ARM
- Further movement from 4->2->1 socket systems as their capabilities expand
- What to do with all the capacity?
- Software defined everything…..



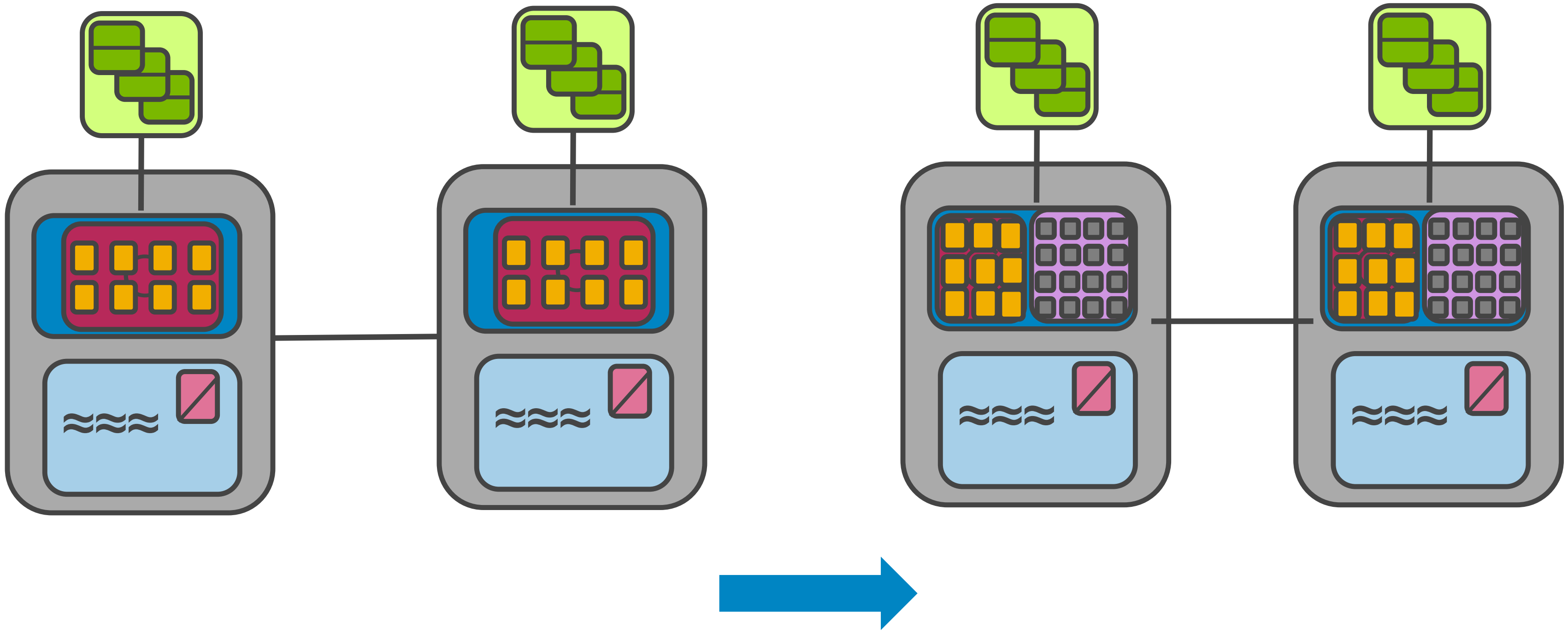
System trend over the years (1)



System trend over the years (2)



The future



System design is being inverted from compute centric to network centric

What levels do we have*?

- Challenge: Sustain performance trajectory without massive increases in cost, power, real estate, and unreliability
- Solutions: <u>No single answer</u>, must **intelligently turn** “Architectural Knobs”

$$\underbrace{\overset{1}{(Freq)} \times \overset{2}{\left(\frac{cores}{socket}\right)} \times \overset{3}{(\#sockets)} \times \overset{4}{\left(\frac{inst\ or\ ops}{core\ \times\ clock}\right)}}_{\text{Hardware performance}} \underbrace{\times\ \overset{5}{(Efficiency)}}_{\substack{\text{What you really get} \\ \text{Software performance}}}$$

*Slide previously from Robert Hormuth

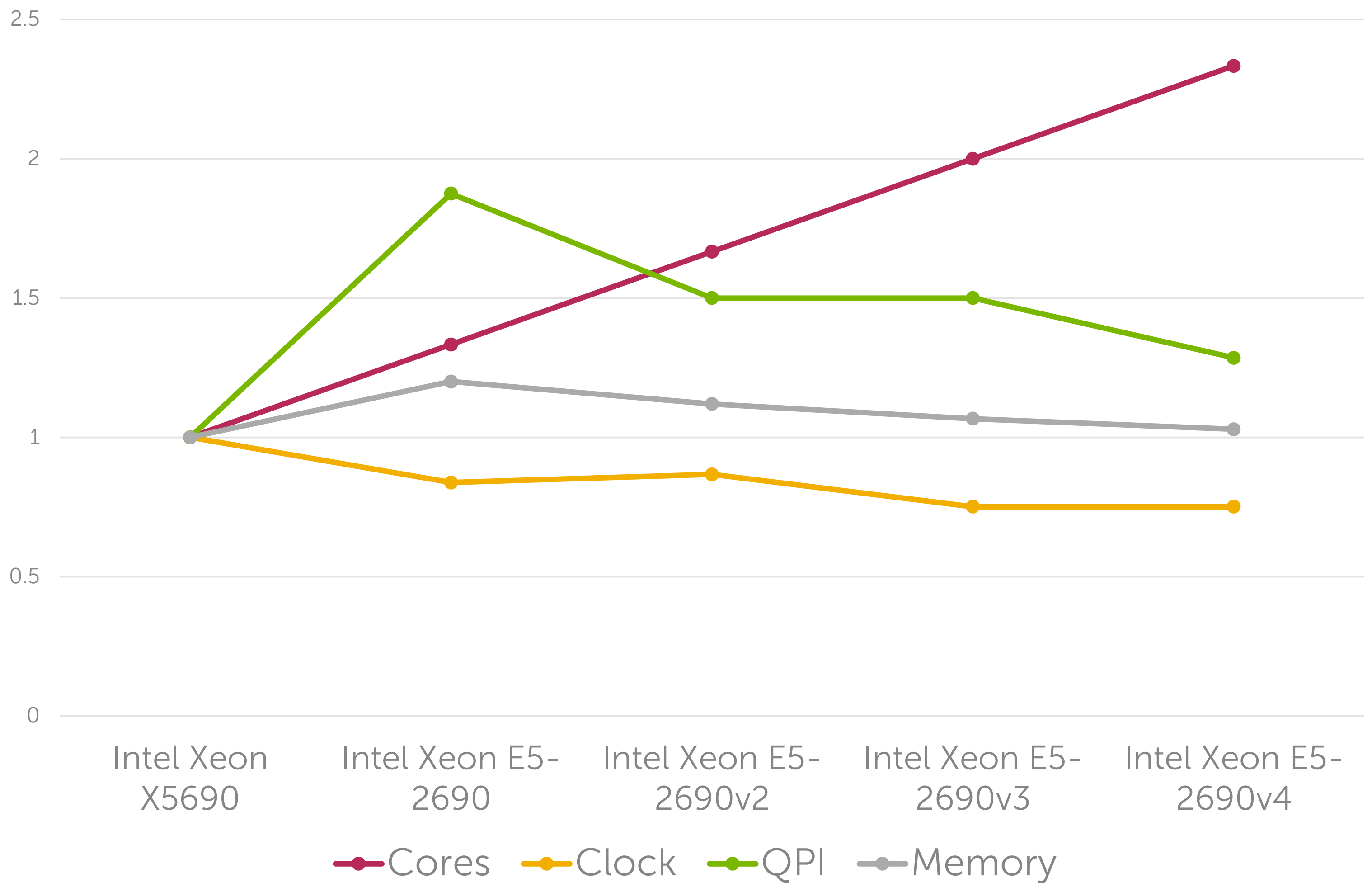
What is Intel telling us?

Intel® Core™ Microarchitecture		Intel® Microarchitecture Codename Nehalem		Intel® Microarchitecture Codename Sandy Bridge		Intel® Microarchitecture Codename Haswell	
Merom	Penryn	Nehalem	Westmere	Sandy Bridge	Ivy Bridge	Haswell	Broadwell
65nm	45nm	45nm	32nm	32nm	22nm	22nm	14nm
New Micro-architecture	New Process Technology	New Micro-architecture	New Process Technology	New Micro-architecture	New Process Technology	New Micro-architecture	New Process Technology
TOCK	*TICK*	*TOCK*	*TICK*	*TOCK*	*TICK*	*TOCK*	*TICK*

New capabilities according to Intel

Intel® Core™ Microarchitecture		Intel® Microarchitecture Codename Nehalem		Intel® Microarchitecture Codename Sandy Bridge		Intel® Microarchitecture Codename Haswell	
Merom	Penryn	Nehalem	Westmere	Sandy Bridge	Ivy Bridge	Haswell	Broadwell
65nm	45nm	45nm	32nm	32nm	22nm	22nm	14nm
New Micro-architecture	New Process Technology	New Micro-architecture	New Process Technology	New Micro-architecture	New Process Technology	New Micro-architecture	New Process Technology
SSE2	*SSSE3*	*SSE4*	*SSE4*	*AVX*	*AVX*	*AVX2*	*AVX2*
2005	*2007*	*2009*	*2011*	*2012*	*2013*	*2014*	*2015*

Meanwhile the bandwidth is suffering



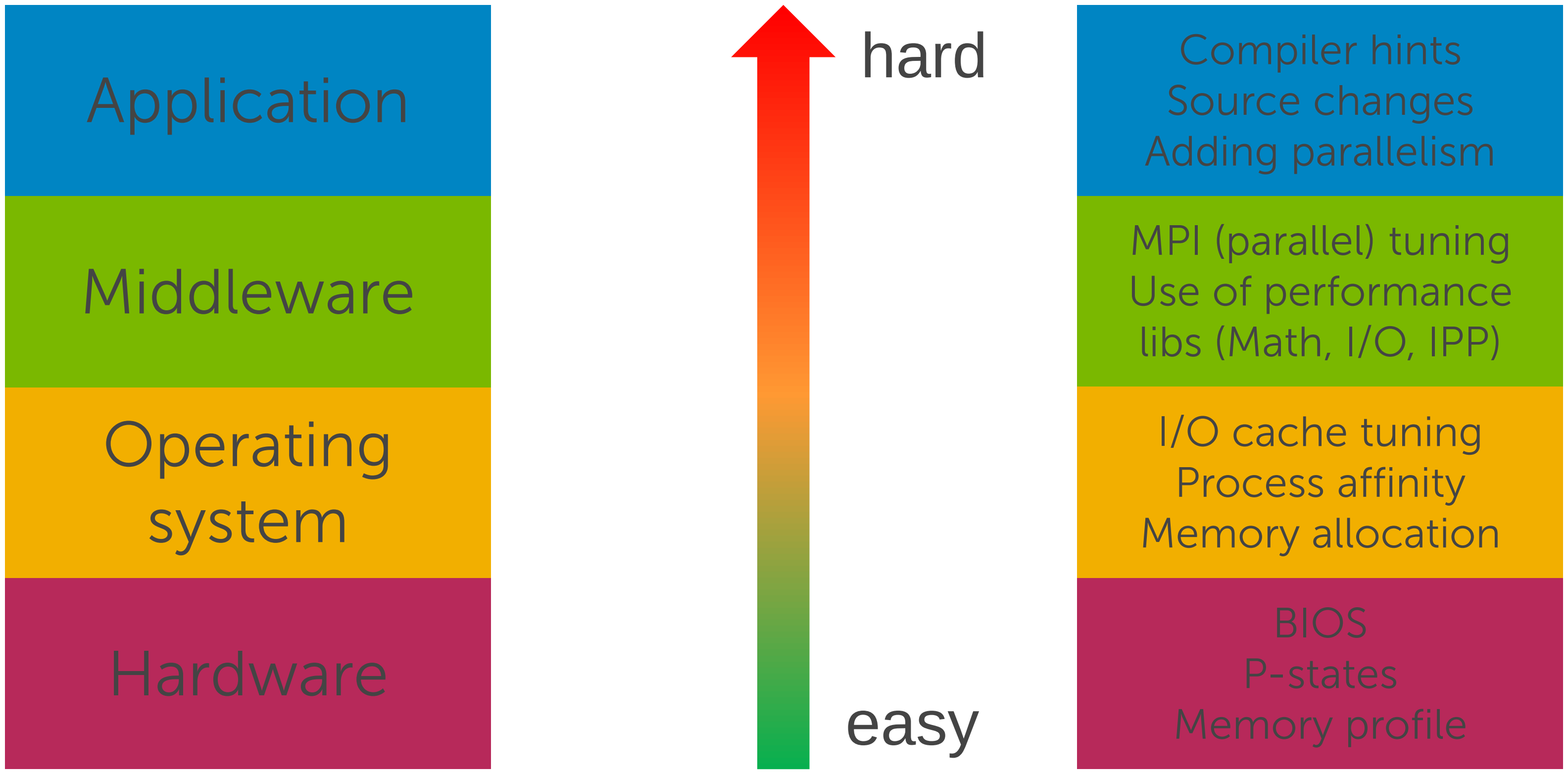
What does Intel do about these trends?

- Providing even more tuning knobs in the hands of the user!

Problem	Westmere	Sandy Bridge	Ivy Bridge	Haswell	Broadwell
QPI bandwidth	No problem	Even better	Two snoop modes	Three snoop modes	Four (!) snoop modes
Memory bandwidth	No problem	Extra memory channel	Larger cache	Extra load/store units	Larger cache
Core frequency	No problem	• More cores • AVX • Better Turbo	• Even more cores • Above TDP Turbo	• Still more cores • AVX2 • Per-core Turbo	• Again even more cores • optimized FMA • Per-core Turbo based on instruction type

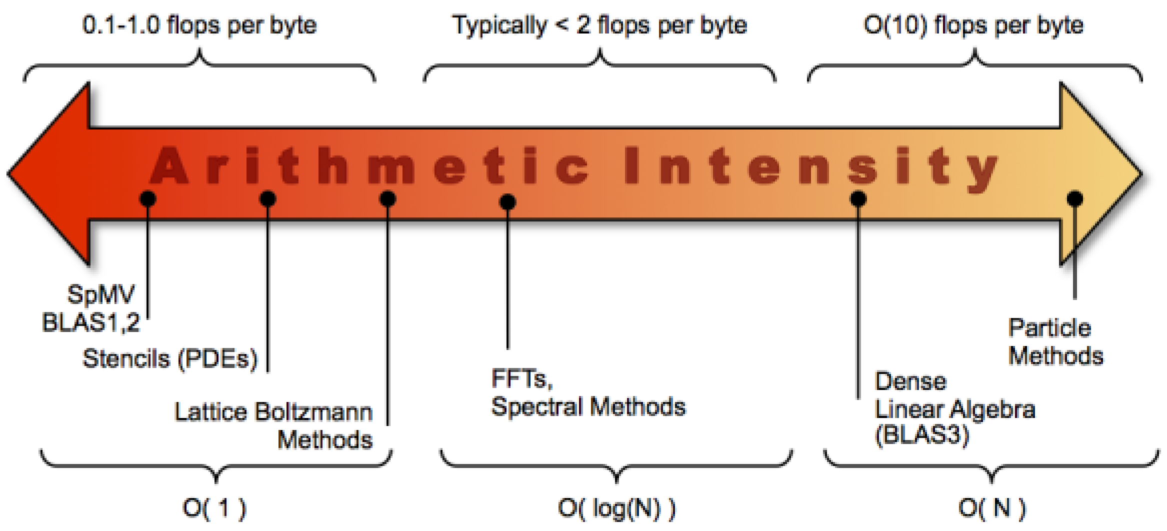
Tuning knobs for performance

Hardware tuning knobs are limited, but there’s far more possible in the software layer



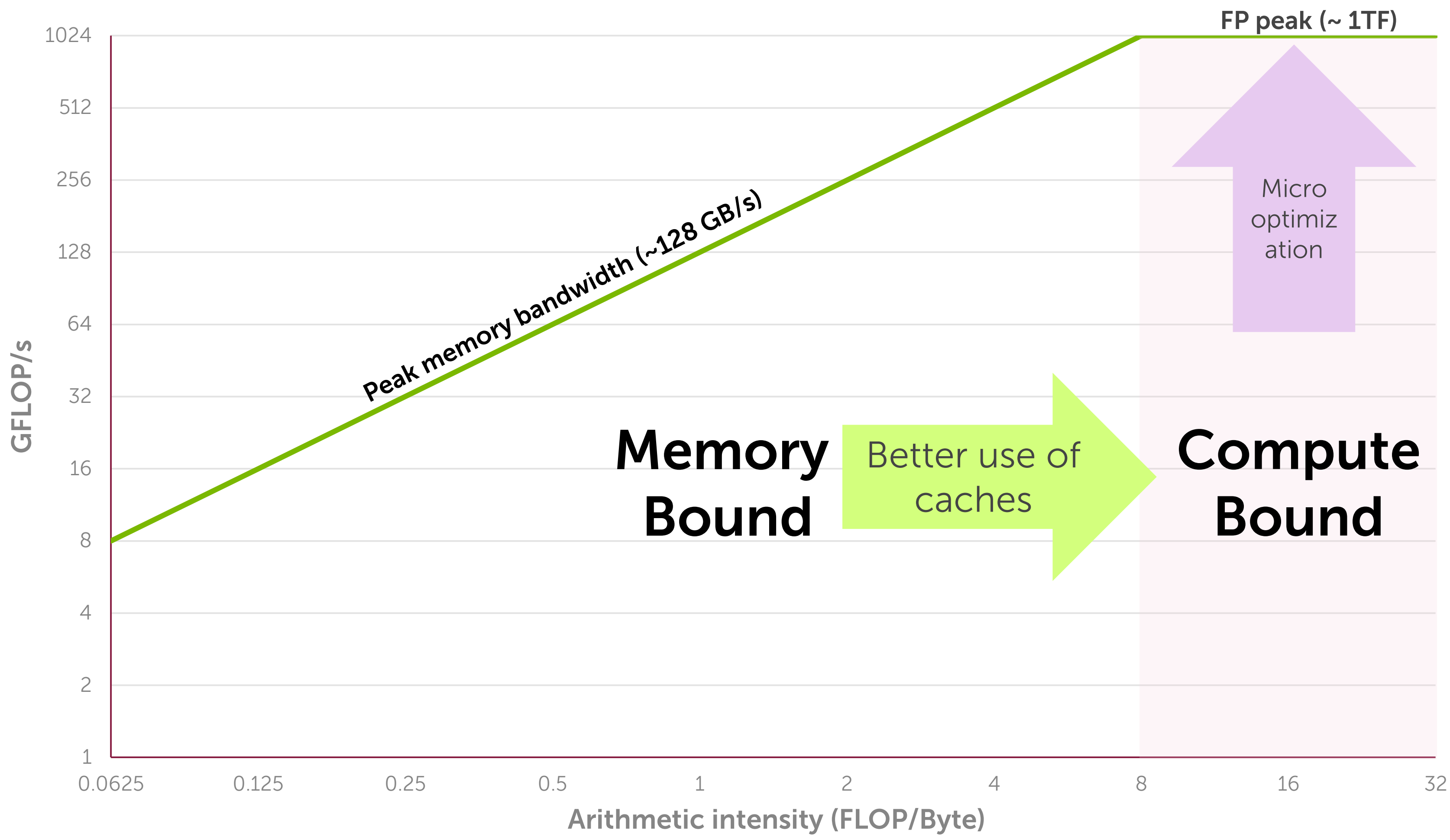
Predicting performance – the roofline model

- Bound system performance as function of peak performance, maximum bandwidth and arithmetic intensity



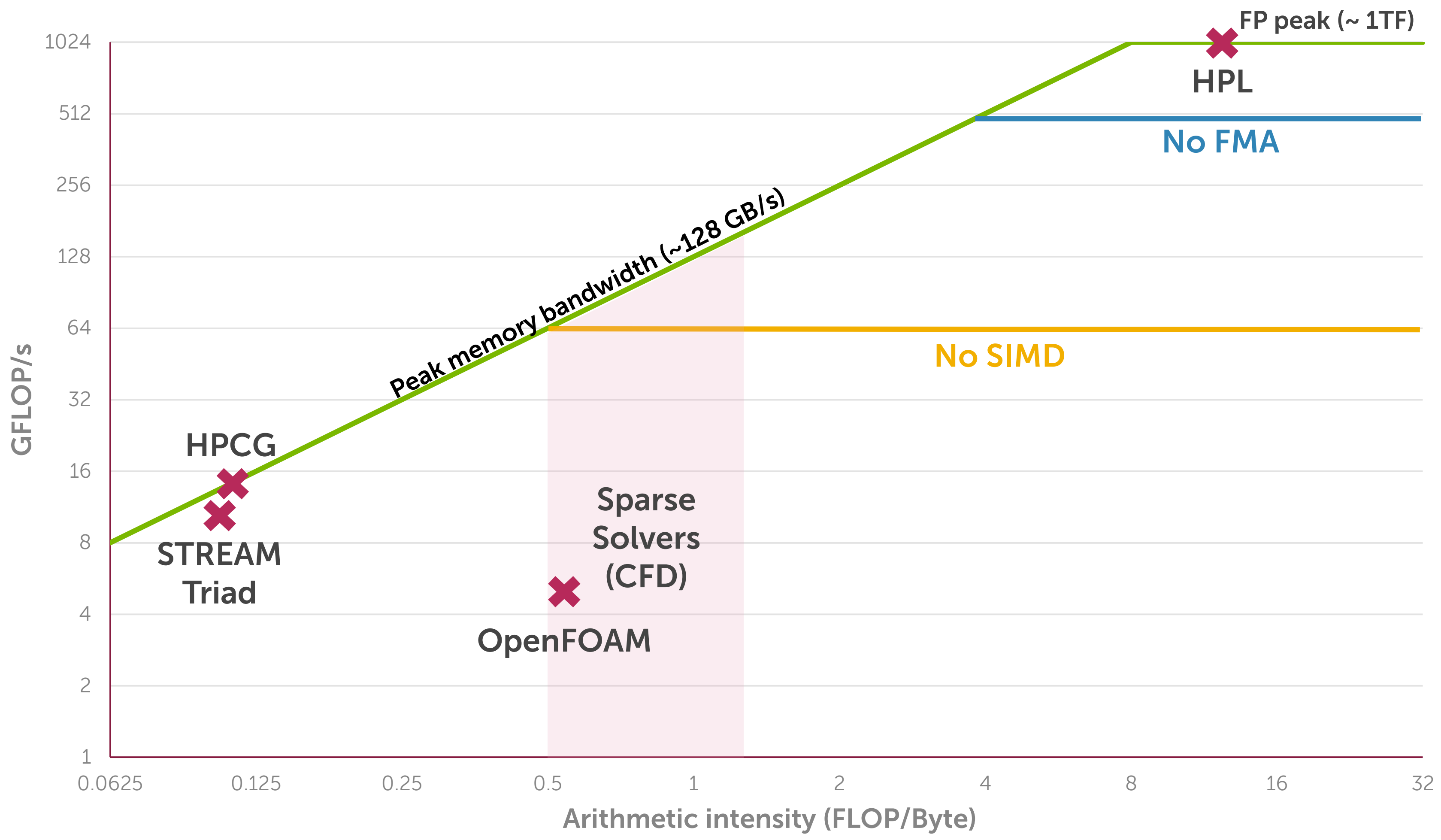
Obtained from: https://crd.lbl.gov/departments/computer-science/PAR/research/roofline/

Roofline model of an E5-2697 v4 processor

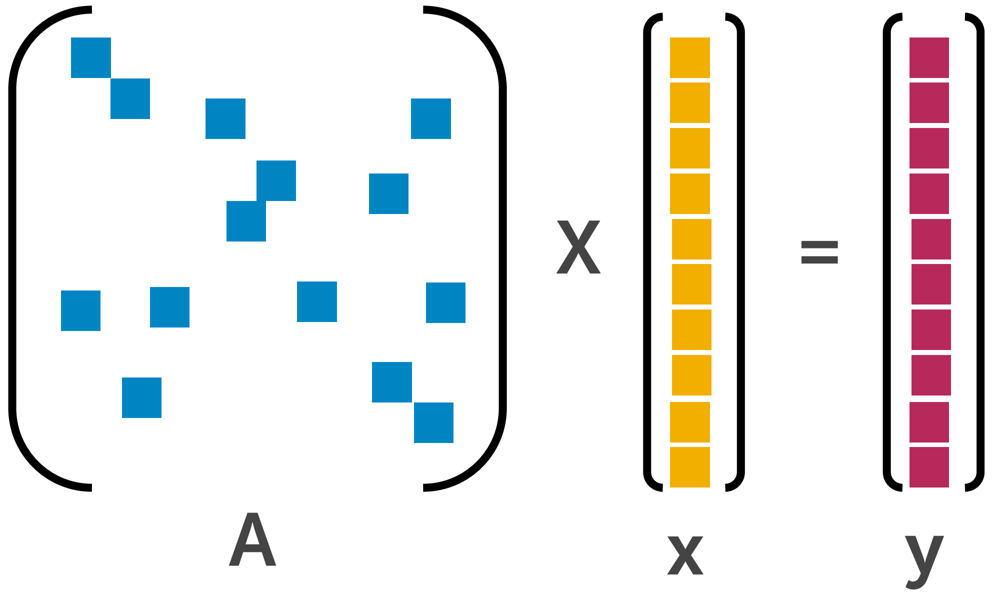


Williams et al.: "Roofline: An Insightful Visual Performance Model", CACM 2009

E5-2697 v4 processor data



Data is becoming sparser (think “Big Data”)



Sparse Matrix “A”

- Most entries are zero
- Hard to exploit SIMD
- Hard to use caches

- This has very low arithmetic density and hence memory bound
- Common in CSM and CFD

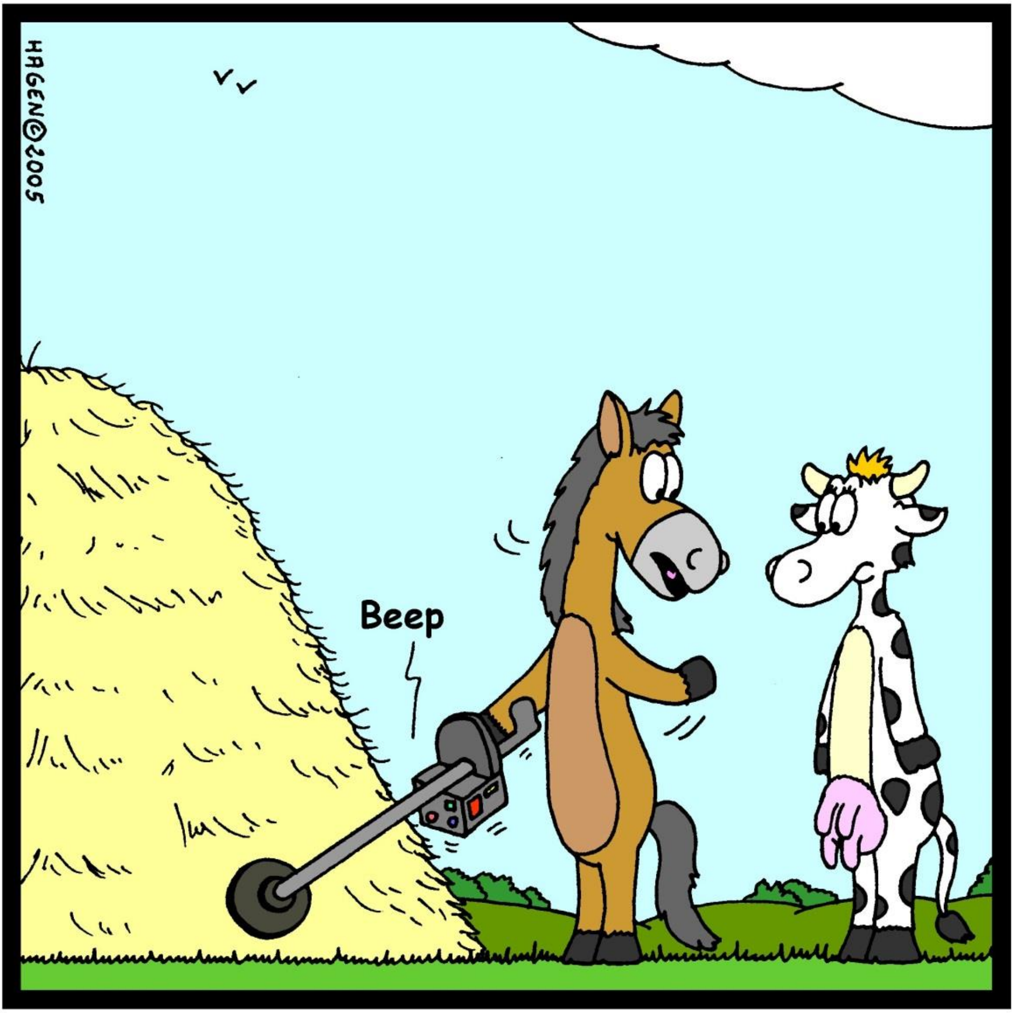
My data used to be here

```
SELECT country_id, country_name
FROM countries;
WHERE region_id = 1;
ORDER BY country_name;
```

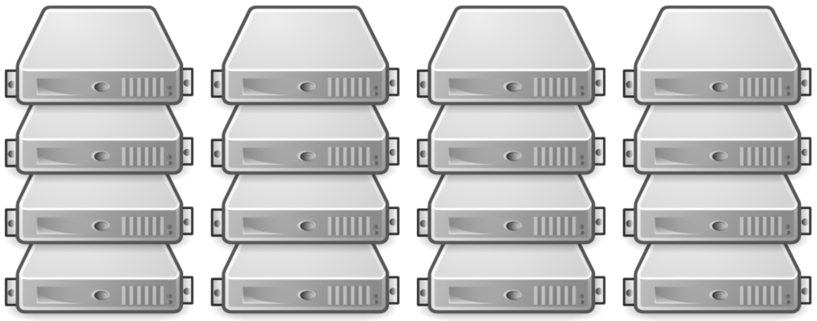
Andorra	Liechtenstein
Austria	Luxembourg
Belgium	Malta
Denmark	Monaco
Finland	Norway
France	Netherlands
Germany	Portugal
Gibraltar	San Marino
Greece	Spain
Iceland	Sweden
Italy	Switzerland
Ireland	United Kingdom

scale vertically – scale up
(bigger box)

But now it is here!



You were right: There's a needle in this haystack...



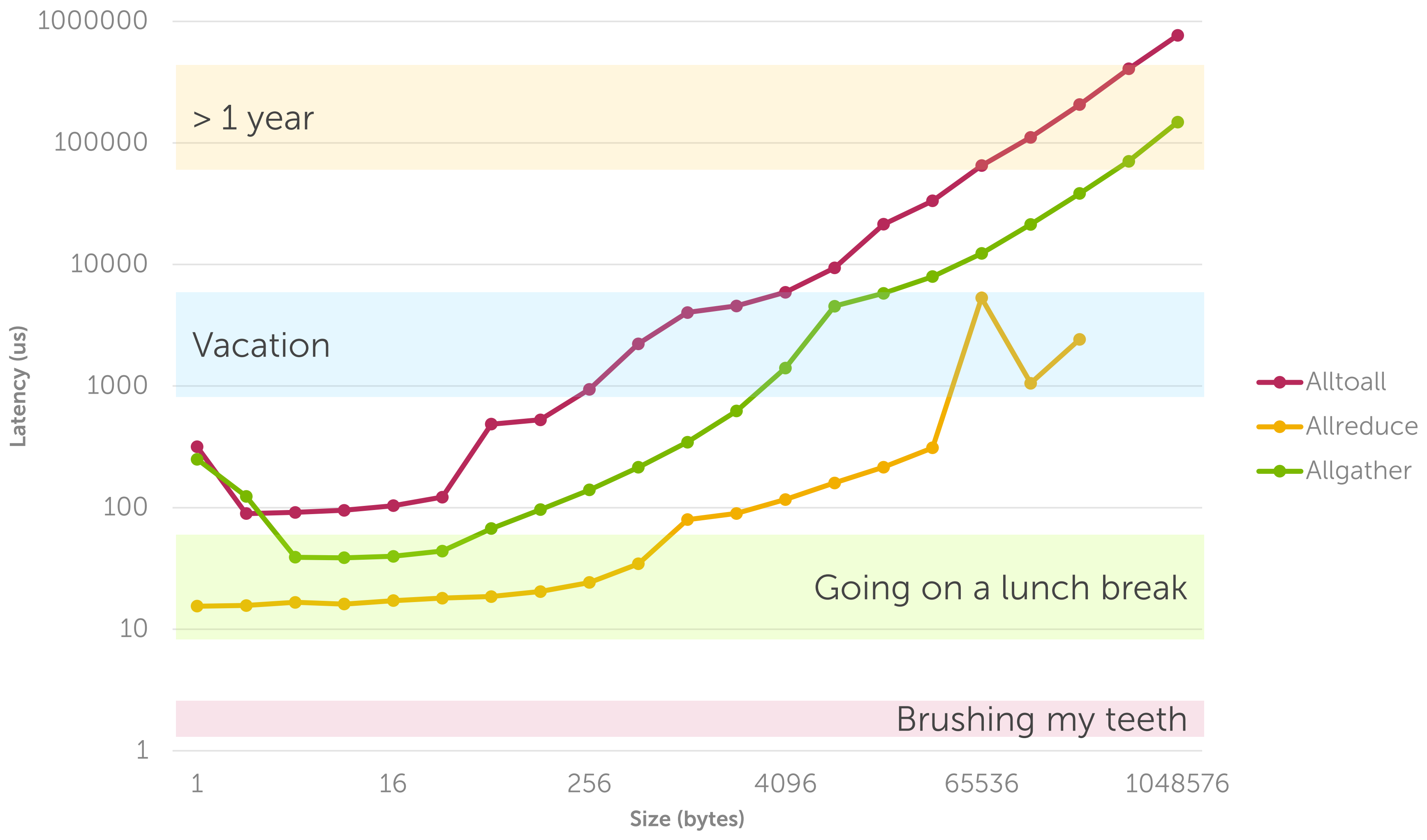
scale horizontally – scale-out
(many small boxes: **cluster**!)

Source: *Hagen Cartoons, http://http://hagencartoons.com/* *Used with permission*

My data is somewhere, but how long does it take to get to me?

Data movement	Latency	Equals to..
L1 cache reference	0.4 ns	One heartbeat
L2 cache reference	5 ns	Long Yawn
L3 cache reference	14 ns	Getting out of bed
Main memory reference	71 ns	Brushing your teeth
MPI ping pong latency	1 us	A run to the grocery store
MPI Allreduce latency (1 kB message)	30 us	FedEx delivery somewhere today
SSD random read	150 us	Weekend
Read 1 MB sequentially from memory	250 us	Holiday weekend
Round trip within data center	0.5 ms	Vacation
Read 1 MB sequentially from SSD	1 ms	Two weeks
Disk seek	10 ms	University semester
Read 1 MB sequentially from disk	20 ms	One year
Send Packet CA->Netherlands->CA	150 ms	Getting a Bachelor's Degree

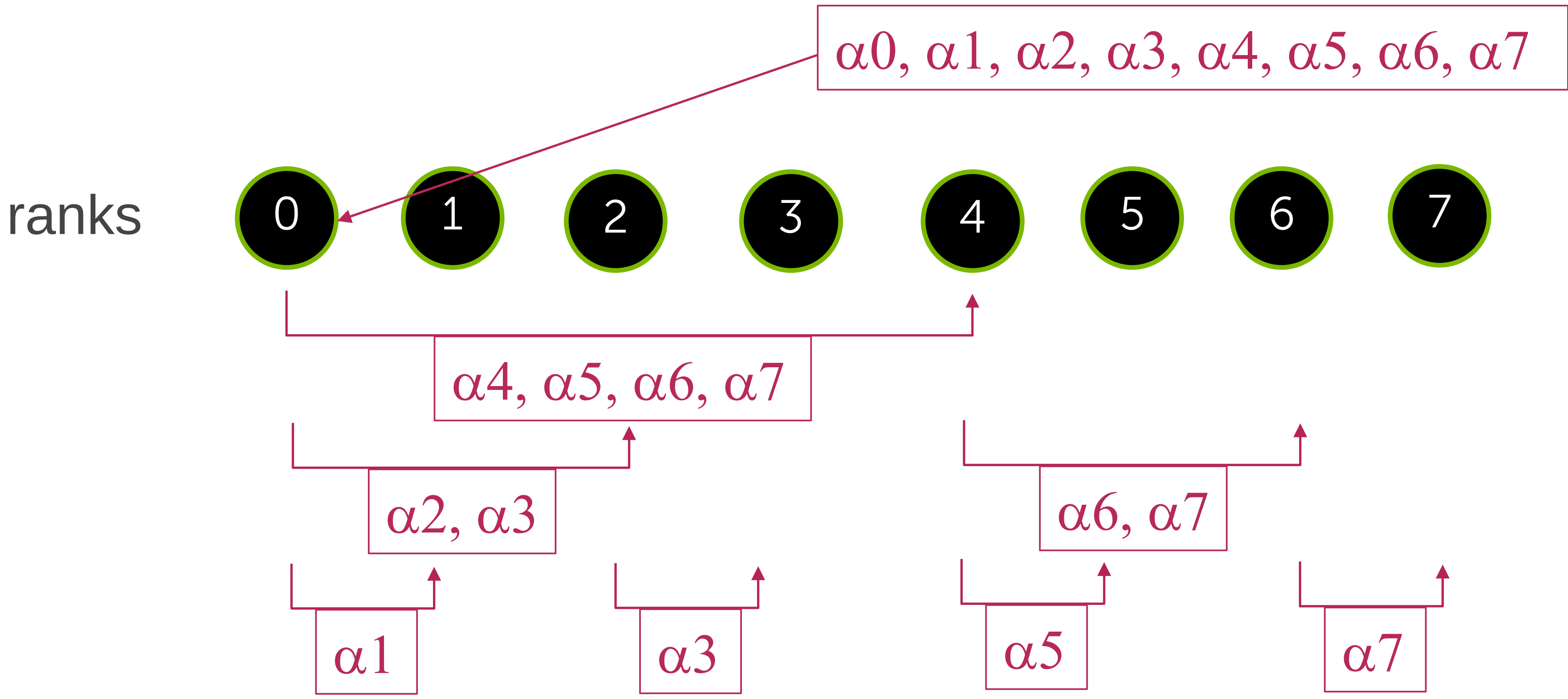
Collective MPI function latency



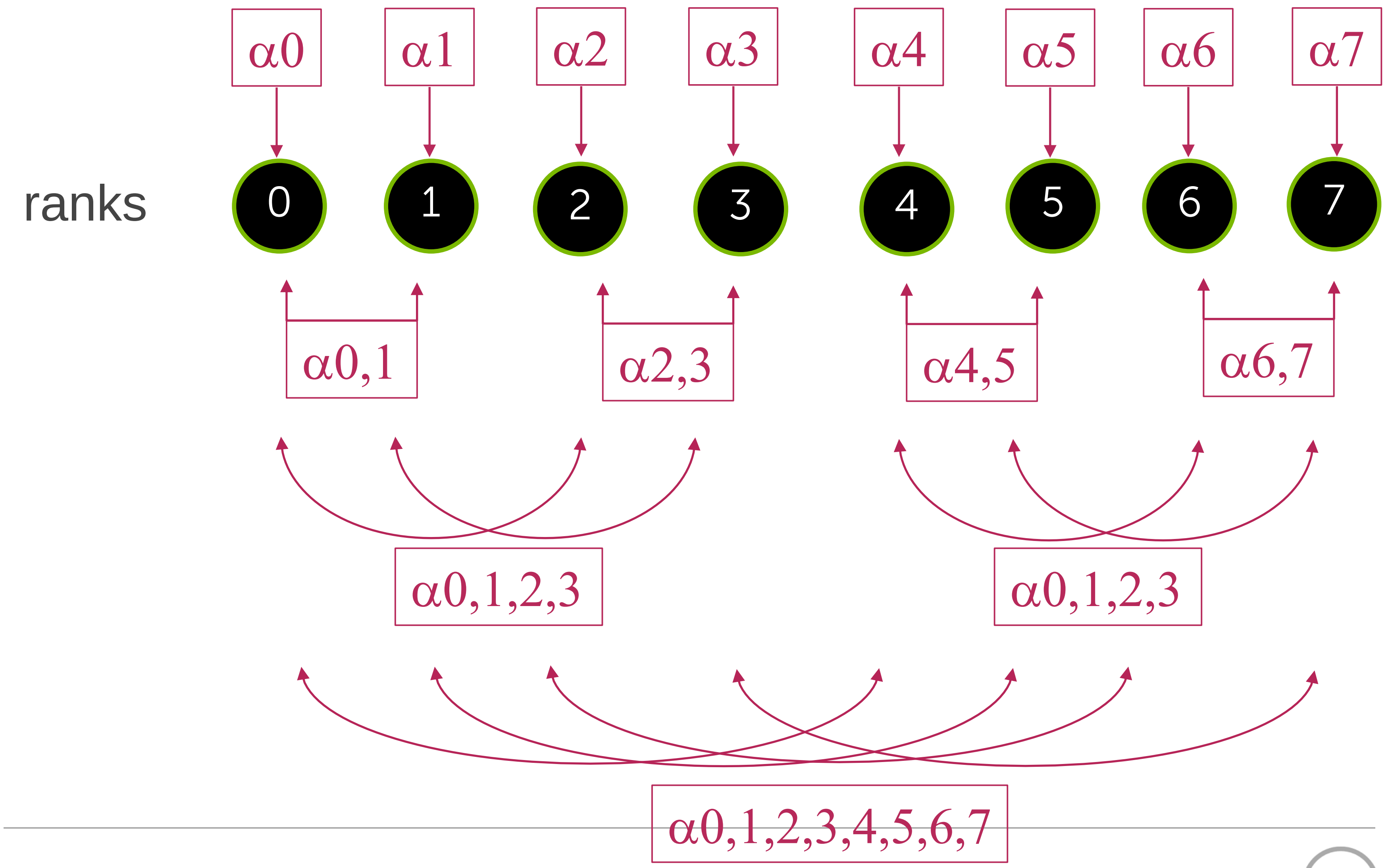
Dell Research Computing

Common network algorithm #1

MPI_Bcast in MPICH – binomial tree



Network algorithm #2: recursive doubling

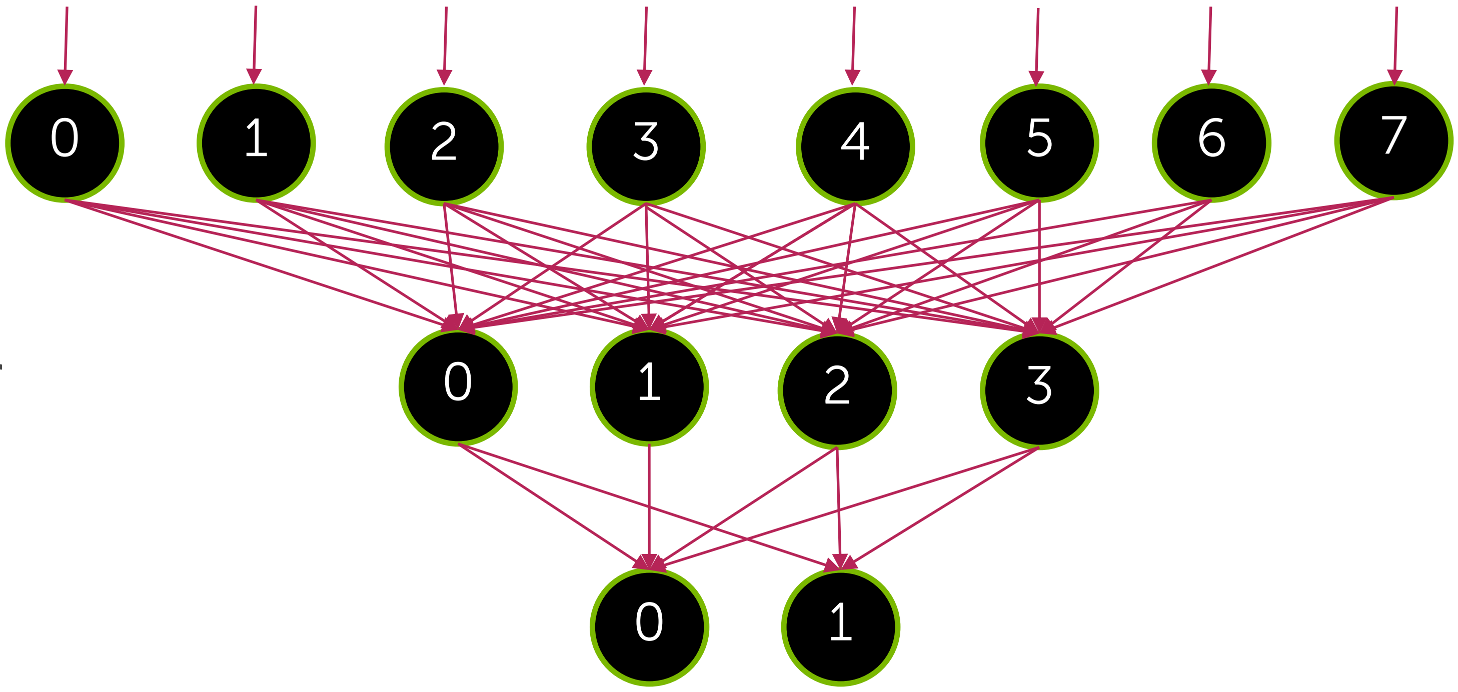


Now compare this with a neural network

Input layer

Hidden layer

Output layer



Data analytics is leveraging HPC architectures

- Data storage keeps growing, but content is becoming sparser
- There is a memory bottleneck to get data to the CPU

- The trend is to replace monolithic systems with small nodes that scale out
- A lot of compute cycles are wasted in communication patterns/algorithms
- Make use of intelligent switching fabrics– **software defined**
- Nowadays the data is closer to the network than the CPU
  - Bring the network closer to the data or the processor closer to the network

The power to do more